Multivariate Stochastic Models of Metocean Fields: Computational Aspects and Applications

A.V. Boukhanovsky

State Oceanographic Institute, St. Petersburg branch, Russia
Institute for High Performance Computing and Data Bases
120 Fontanka ,St. Petersburg, 198005, Russia
E-mail: avb@fn.csa.ru

1. Introduction

Metocean data fields (atmospheric pressure, wind speed, sea surface and air temperature, sea waves etc.) are multivariate and multidimensional, i.e. have a complex spatial and temporal variability. Only 10-20 years back the environmental databases wholly consisted of the time series (ship observation, sea and coastal monitoring stations, automatic probes and buoys, satellites) in fixed points of spatial regions. For processing of such data the different kinds of software are developed, e.g. [11]. They interprets the information in terms of random values (RV) or time series (TS) models only.

Development of environmental hydrodynamical models and use them for data assimilation and reanalysis [14], has allowed to create global information arrays of metocean fields in points of a regular spatial-temporal grid. So, the results of meteorological fields reanalysis may be used as source for hydrodynamic simulation of metocean (generally, hydrologic) fields, e.g., sea waves fields [8], water temperature and salinity fields [12] et al. This way allow to obtain the ensemble of metocean data fields in a regular grid points with certain temporal step. Hence, for processing and generalization of such data the model of a nonstationary inhomogeneous spatial-temporal random field (RF) must be considers.

Due to the high dimension of data in set of grid points, the multivariate statistical analysis (MSA) are applied to their processing. The goal of MSA is solution of three global problems - reduction of dimensionality (RD), detection of dependence (DD), and detection of heterogeneity (DH) of the random data [1]. Its operates by the canonical variables (principal components, canonical correlations, factor loadings), includes regression, discriminant analyses, and analysis of variance and covariance, classification, clustering and multidimensional scaling.

The main complexity of direct application of classical MSA to analysis of metocean spatial–temporal fields is connected with the fact, that all the procedures are developed for model of multivariate RV only [1,2,15,19]. Some of them are generalized for model of multivariate time series [5]. But there are two problems in the generalization of these procedures to RF model:

- ***Various physical nature of metocean fields.*** Hence, the various mathematical abstract objects may be used for their statistical description. E.g. the field of

P.M.A. Sloot et al. (Eds.): ICCS 2002, LNCS 2329, pp. 216–225, 2002.

atmospheric pressure is scalar function, and the simplest moments (mathematical expectation and variance) are scalar values too. The wind speed field is Euclidean (geometric) vector function. The mathematical expectation of such kind of data is Euclidean vector too, and variance is dyadic tensor. The field of water parameters (temperature, salinity, oxygen) is the affine (algebraic) vector function with affine vector of mathematical expectation and matrix of variance. Hence, it is necessary previously to introduce the basic algebraic and statistical operation for each kind of abstract objects. General discussion about this is in the paper [6].

- ***Requirements to computational aspects of MSA RF procedures.*** For traditional MSA RV with sufficiently small number of variables, the methods of linear algebra are developed well [10]. But for random fields of high dimension, especially in a small spatial and temporal grids, some of results by these methods became ill–conditioned due to strong spatial and temporal connectivity of data. Hence, it is necessary to develop special computational tools for MSA RF on the base of functional analysis in infinite–dimensional spaces of functions [3].

The goals of this paper are:

- To demonstrate the general approach for MSA RF in arbitrary functional space (scalar, Euclidean or affine vector) considering three main problems: RD, DD, DH.
- To synthesize the stochastic models on the base of MSA RV result for further ensemble simulation for investigation of non–observable rare environmental events.

2. General Approach

Let us consider the infinite–dimensional space of functions **H** with the operations of addition $(\mathbf{f+g})\in\mathbf{H}$, multiplication $(\mathbf{f}\circ\mathbf{g})\in\mathbf{H}$, and scalar product $(\mathbf{g,f})\in\mathbf{R}$ of elements **f, g**. Concept of scalar product is obvious only to scalar values. For Euclidean and affine vectors it generalizes concept of scalar product both in discrete space and in continuous space. If we consider the functional spaces of metocean events as Hilbert [3], in each of them any element η can be presented as an infinite converging series on some system of basic elements (scalar, Euclidean of affine functions) $\{\varphi_k\}$:

$$\eta = \sum_k c_k \phi_k \,. \tag{1}$$

The back transformation

$$c_k = (\eta, \phi_k) \tag{2}$$

defines an isomorphism between functional space **H** and discrete space $\mathbf{C}$, $c_k\in\mathbf{C}$.

Due to linearity of (1,2), the mathematical expectation and variance of η may be expressed from the moments of coefficients $\mathbf{c_k}$:

$$\mathbf{m_\eta = \sum_k m_k \phi_k} \ , \tag{3}$$

$$\mathbf{K_\eta = \sum_k \sum_p k_{kp} [\phi_k \circ \phi_p]} \ . \tag{4}$$

Here $\mathbf{m_k}$ – scalar mathematical expectations of $\mathbf{c_k}$, $\mathbf{k_{kp}}$ – covariances between $\mathbf{c_k}$, $\mathbf{c_p}$.

Full system of orthogonal functions $\{\phi_k\}$ may be obtained by means of Gram–Shmidt [3] orthogonalization procedure in accordance with of equation

$$(\phi_k, \phi_s) = \delta_{ks} N_k \ , \tag{5}$$

where N_k is a norm. Even for Euclidean or affine vector fields it is complicate to obtain such functions in convenient form.

Let us consider optimal basis $\{\varphi_k\}$ from all $\{\phi_k\} \in \mathbf{H}$. Following [17], this basis is generates by the equation

$$(K_\eta, \varphi_k) = D_k \varphi_k \tag{6}$$

in the certain functional space **H.** System $\{\varphi_k\}$ is orthogonal, but coefficients $\mathbf{c_k}$ are independent random variables with variances $\mathbf{D_k}$. For all other choices of $\{\phi_k\}$ coefficients $\mathbf{c_k}$ are correlated due to (4).

The main advantage of model (1) is passage from random field η (scalar, Euclidean or affine vector multivariate function) to set of scalar coefficients or scalar time series $\{\mathbf{c_k}\}$. Hence, for different kinds of data the model of RV is valid for $\{\mathbf{c_k}\}$. Below let us consider some applications of this approach for analysis and synthesis of various metocean fields.

3. Applications to analysis

3.1. Reduction of dimensionality [6]**.** Let us consider the scalar fields of sea level pressure (SLP) $\zeta(\vec{\mathbf{r}}, t)$ and wind speed (VS) at the level 10 (m) $\vec{\mathbf{V}}(\vec{\mathbf{r}}, t)$ by the reanalysis [13] data array. Here $\vec{\mathbf{r}} = (x,y)$ are the spatial coordinates, $\mathbf{t}$ is time. Traditionally for RD the expansion on empirical orthogonal basis (EOF [14]) are uses. For scalar fields of SLP the EOFs are the eigenfunctions of correlation kernel, and equation (6) became to Fredholm II integral equation [17]:

$$\int_{\langle \mathbf{R} \rangle} \mathbf{K}(\vec{\mathbf{r}}, \vec{\mathbf{r}}_1) \varphi_k(\vec{\mathbf{r}}_1) d\vec{\mathbf{r}}_1 = \lambda \varphi(\vec{\mathbf{r}}) \ . \tag{7}$$

Spectrum $\{\lambda_k\}$ of such kernel is discrete. Application of quadrature methods for (7) solution results in matrix representation without avoiding multicollinearity of the

problem. Therefore let us use projective (variational) methods [11] for obtaining orthogonal basis. The idea of this methods is using of expansion (1) for representation of each EOF in (7). If we substitute (3) in (7), the equation (7) transform to algebraic eigenvalue problem. Occasionally such method allows to obtain analytical solution for some types of modeling representations of autocorrelation function $\mathbf{K_\eta(\bullet)}$ of non-homogeneous field [5].

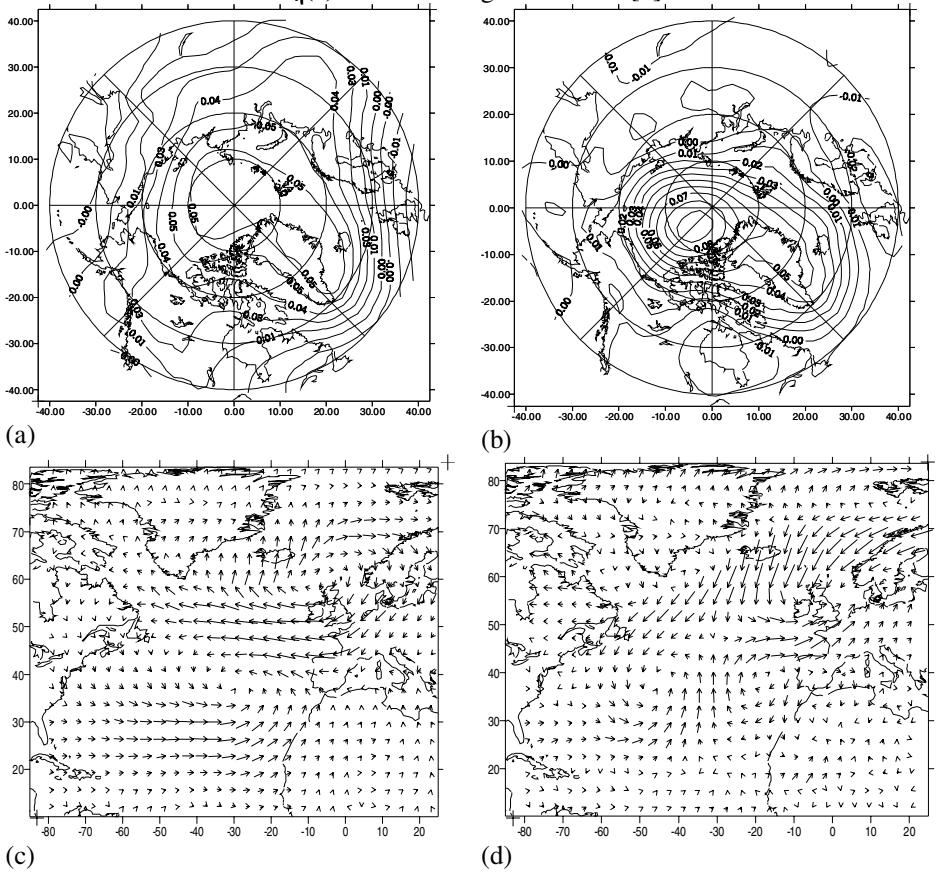

Fig. 1. First EOFs of monthly–mean SLP (North Hemisphere) and VS (North Atlantic). (a,c)–January, (b,d)–July

When we generalize (1) on Euclidean vector field $\vec{\mathbf{V}}(\vec{\mathbf{r}},t)$ in accordance with the rules of vector sums and dyadic (external) products, the equation (6) transforms in two homogeneous Fredholm II equations:

$$\int \mathbf{K_{uu}}(\vec{r}_1,\vec{r}_2)\varphi(\vec{r}_2)d\vec{r}_2 + \int \mathbf{K_{uv}}(\vec{r}_1,\vec{r}_2)\psi(\vec{r}_2)d\vec{r}_2 = \lambda\varphi(\vec{r}_1), \tag{8}$$

$$\int \mathbf{K_{vu}}(\vec{r}_1,\vec{r}_2)\varphi(\vec{r}_2)d\vec{r}_2 + \int \mathbf{K_{vv}}(\vec{r}_1,\vec{r}_2)\psi(\vec{r}_2)d\vec{r}_2 = \lambda\psi(\vec{r}_1),$$

with respect to components $\vec{\Psi} = (\varphi, \psi)$. Here $K_{\bullet\bullet}$ – correlation functions between components of Euclidean vector. Using of (8) allows to simplify the interpretation of vector EOFs. E.g., in fig. 1 the first EOFs of monthly–mean SLP (a,b) and VS (c,d) are shown for January and July. It is clearly seen the great zones of wind speed variability near Iceland and Azores.

First ten EOF's of monthly–mean SLP determined more than 80–90% of general variability (from month to month), and first ten EOF's of WS – more than 70%.

2. Detection of dependence [7]. Let us consider the long–term dependence between iceness (the area of the sea, covered by ice) of Barents Sea, and air temperature spatial–temporal fields (below – AT). Take in account joint spatial and temporal variability, we define the iceness time series $\zeta(t)$ in terms of multivariate dynamical system [4]:

$$\zeta(t) = \int_{<R>} \int_{0}^{\infty} h(\vec{r}, t-\tau)\eta(\vec{r}, \tau)d\tau d\vec{r} + \varepsilon(t). \tag{9}$$

Here $\mathbf{h(\bullet)}$ is transfer functions between input η (AT) and output ζ (iceness), $\varepsilon(t)$ – is lag random function.

The values of AT in nearest points are strongly correlated. So, the problem of multicollinearity is observes, and nonparametrical estimation of $\mathbf{h(\bullet)}$ in (9) is non–correct. For increasing of conditionality of equations, born by (9), let us consider the AT fields as an expansion (1) by EOFs from (6,7). The first EOF approximated more 90% of total variability.

The second step of model simplification is using the orthogonal decomposition in time domain. While the series of iceness (see fig. 2) are clearly cyclic, the model of periodically correlated stochastic process (PCSP) are used for it representation. So, the time series may be expanded by trigonometric basis. Generally we obtaining bi–orthogonal expansion for AT fields

$$\eta(\vec{r}, t) = \sum_{n} a_n(t)\psi_n(\vec{r}) = \sum_{n}\sum_{m} b_{nm}^{(c)}\psi_n(\vec{r})\cos(\omega_m t) + b_{nm}^{(s)}\psi_n(\vec{r})\sin(\omega_m t) \tag{10}$$

and expansion for iceness time series

$$\zeta(t) = \sum_{m} \beta_m^{(c)} \cos(\omega_m t) + \beta_m^{(s)} \sin(\omega_m t). \tag{11}$$

This model allows to transfer from model of correlated RFs to model of depended RV $\left\{b_k^{(s)}, b_j^{(c)}, \beta_m^{(s)}, \beta_n^{(c)}\right\}$, so the procedure of dependence detecting is simplified. The joint correlation matrix $\mathbf{K}\left[b_k^{(s)}, b_j^{(c)}, \beta_m^{(s)}, \beta_n^{(c)}\right]$ is rarefied, but between several coefficients the dependence is high. If we calculate the canonical correlations [15] between $\left\{b_k^{(s)}, b_j^{(c)}\right\}$ and $\left\{\beta_m^{(s)}, \beta_n^{(c)}\right\}$, we obtain λ_1=**0.8** and λ_2=**0.6**, so, the general dependence between considered factors is sufficiently high. In the fig. 2 the time

series of iceness from 1960 to 1980 are shown by observations and model simulation by (9–11). It is clearly seen the good agreement between to graphs.

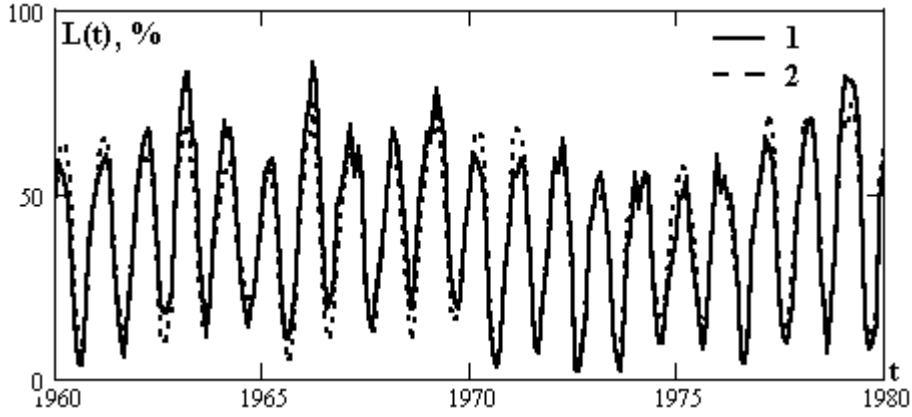

Fig. 2. Time series of iceness (%) of the Barents sea. 1 – observations, 2 – simulation (9–11).

3. Detection of heterogeneity [8]. Let us consider the example of typization of water mass of Baltic sea on the base of long–term observations on four International monitoring stations BY–2, BY–5, BY–15, BY–28, see fig. 3a. The water temperature **T**, salinity **S**, conditional density **ρ** and dissolved oxygen concentration $\mathbf{O_2}$ are jointly used for typization. All the data obtained each month from 1960 to 1980, in standard depths: z=0(10)100(25)… meters.

The generalization of assimilated data in terms of mathematical expectations and variances of initial data demands to operate with the grid function of mathematical expectation in 8832 points, and correlation matrix with more than 39·106 independent elements.

For decreasing of dimensionality the bi–ortogonal expansion of (1) on the basis $\{\varphi_k(z),\psi_s(t)\}$ for each oceanographic value $(\mathbf{T,S,\rho,O_2})$ is considered

$$\zeta(z,t) = \sum_k a_k(t)\varphi_k(z) = \sum_k \sum_s b_{ks}\varphi_k(z)\psi_s(t) = \sum_s \eta_s(z)\psi_s(t)\,. \qquad (12)$$

In paper [8] shown, that the spatial basis $\varphi_k(z)$ is Chebyshev polynomials, and temporal one $\psi_s(t)$ – is trigonometric. So, the expression (12) allows to transfer from model of vertical–inhomogeneous RF $\zeta(z,t)$ to system of RV $\{b_{ks}\}$ for $\mathbf{T,S,\rho,O_2}$ independently. Due to the physical dependence between water mass parameters, the values $\{b_{ks}\}$ are correlated. Let us use its for typization in terms of discriminant variables. One of general procedures of discriminant analysis is obtaining of canonical discriminant functions (KDF) f_{pm}, as follows linear combination [15]:

$$f_{pm} = u_0 + \sum_k \sum_s u_{ks} b_{ks}^{(pm)} \qquad (13)$$

Here $\mathbf{b}_{ks}^{(pm)}$ are coefficients of expansion (12) for monitoring station **p** in a year **m**, $\mathbf{u_0}$, $\mathbf{u_{ks}}$ are the coefficients, so, that the distinctions of mathematical expectations of two each classes became maximal. Values of $\mathbf{f_{pm}}$ are not correlated.

For obtaining of CDF the matrix equation are used:

$$\mathbf{B\nu = \lambda W\nu} \qquad (14)$$

where $\mathbf{u_i = \nu_i\sqrt{n-g}}$, $\mathbf{u_0 = -\sum_{i=1}^{n} u_i x_i}$. Here **B** is matrix of correlations between classes, **W** – matrix of correlations inside the classes (for each station BY), $\boldsymbol{\nu}$ – eigenvectors and $\boldsymbol{\lambda}$ – eigenvalues respectively of matrix $\mathbf{W^{-1}B}$. The first CDF explain approximately 50% of general distinctions, and second – 45%. So, only two first CDF may be used for typization of water masses. In the fig. 3 in plane of first two CDF $(\mathbf{f_1,f_2})$ the classes of observations are shown. It is clearly seen, that there are three separated classes: Baltic sea near the Darss (BY–2), Southern Baltic (BY–5) and Central Baltic (BY–15,28).

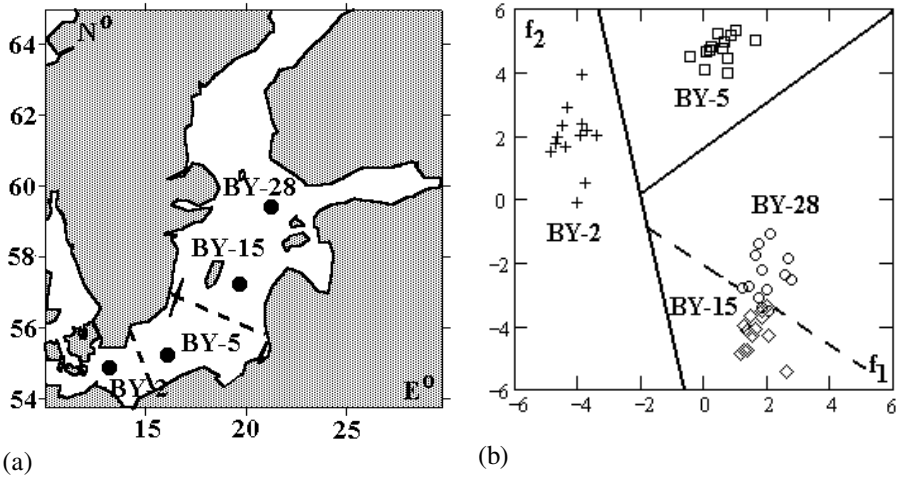

Fig. 3. Fig. 3. Typization of water masses of Baltic sea. (a) – Positions of International monitoring stations (b) – Data of four monitoring stations in first two CDF $(\mathbf{f_1,f_2})$ plane.

4. Application to ensemble simulation and extreme analysis

Representation (1) allows to synthesize the model ensemble of metocean fields on the base of MSA RF results by means of stochastic models. These models are used for interval estimation, testing of hypotheses about latent properties of phenomena and investigation of rare non–observable situations, e.g. extreme values of metocean fields once T years.

For identification of stochastic model let us consider (1) in terms of factor analysis [1]:

$$\eta = \sum_{k} c_k \phi_k + \varepsilon . \qquad (15)$$

Here c_k are the factors, basis elements ϕ_k – became the factor loadings, and ε is the specific factor. If all the factors obtained by means (6), this method called as "principal factor method" for scalar variables.

Model (15) allows generate the model ensemble by means of Monte–Carlo approach on the base of variances of coefficients $\mathbf{c_k}$, and specific factor $\boldsymbol{\varepsilon}$ [20]. For example, in the fig. 4 the directional annual WS extremes in North Atlantic (55N,30W) are shown. These values are obtained by means of direct estimation on simulated ensemble [19] by the model, based on EOF's from fig.1. It is seen, that strongest direction is West (28 m/s), and weakest – is North–East.

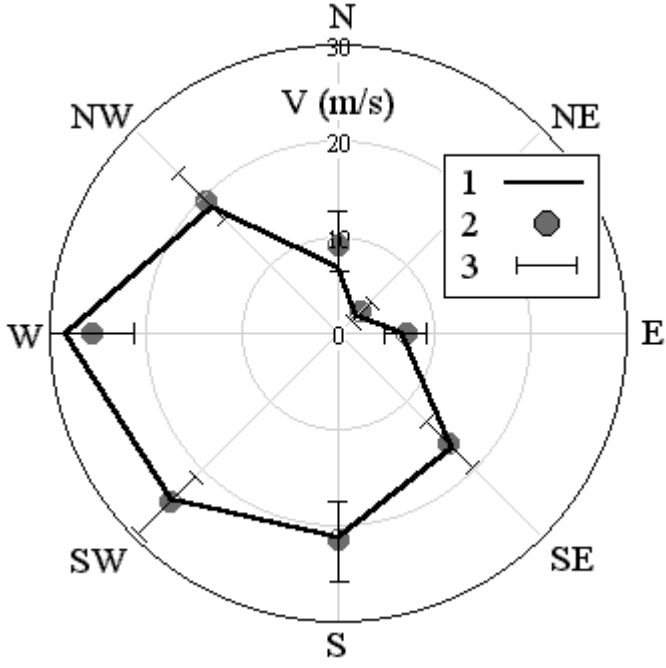

Fig. 4. Estimation of annual WS directional extremes in North Atlantic (55N,30E). 1 – simulation, 2 – estimation on reanalysis data, 3 – 95% confidence intervals

In the same fig. the values of annual extremes on reanalysis [13] data are shown too. The 95% confidence intervals covered both simulated and reanalysis estimates. Thus, it shown the good agreement between stochastic model and data.

5. Discussion and conclusions

The above presented examples shown, that the using of orthogonal expansion (1) for representation of multivariate and multidimensional spatial–temporal metocean fields allows to obtain some advantages in comparison with classical matrix methods of MSA RV, e.g.:

- Simplification of probabilistic model (from non–scalar RF to system of scalar RV);
- Possibility to obtain the analytical solutions – a step for complex mathematical investigation of physical phenomena;
- Avoiding of quantization problem (methods of interpolation between greed points and objective analysis are not required);
- Decreasing of data dimensionality (in example 3, the initial data correlation matrix consists of more than $39 \cdot 10^6$ independent elements, and reduced matrix in (14) – only 153 elements);
- 100% parallelization by data due to independent computation of each values of coefficients (2) by means of scalar product;
- Improvement of matrix conditionality. The orthogonal expansion (1) in matrix terms may be consider as singular value decomposition (SVD) [10] of correlation matrix by means of orthogonal matrixes of special type. So, the result of such SVD is matrix of smaller rank with the smaller conditional number. So, the stability of computations with such matrix is higher.

It is necessary to indicate some shortcuts of the above mentioned approach. One of them is based in the correct choice of type of formal orthogonal basis. In some cases, as temporal rhythms, this choice is obvious, but for description of spatial patterns the problem has no unique solution. With this fact the weak convergence of the expansions may be connected.

Despite of this, the orthogonal expansions (1) in functional spaces remains the power tool of multivariate statistics of spatial–temporal fields and may be used for all the problems: reduction of dimensionality, detection of dependence and detection of heterogeneity. Using of this tool is not be based only on formal statistic approach, but the specifics of physical phenomena, initial database and hindcast tools must take into account.

Acknowledgement

This research is supported by INTAS grant Open 1999 N666.

References

1. Anderson T.W. An introduction to multivariate statistical analysis. John Wiley, NY, 1948.

2. Bartlett M.S. Multivariate analysis. J. Roy. Stat. Soc. Suppl. 9(B), 1947, 176–197.
3. Balacrishnan, Applied functional analysis. NewYork, John Wiley, 1980
4. Bendat J.S., Piersol A.G. Random data. Analysis and measurement procedures. John Wiley, NY, 1989.
5. Brilinger D. Time series. Data analysis and theory. Holt, Renehart and Winston, Inc., New York, 1975.
6. Boukhanovsky A.V., Degtyarev A.B., Rozhkov V.A. Peculiarities of computer simulation and statistical representation of time–spatial metocean fields. LNCS #2073, Springer–Verlag, 2001, pp.463–472.
7. Boukhanovsky A.V., Ivanov N.E., Makarova A.V. Probabilistic analysis of spatial–temporal metocean fields. Proc. of Reg. Conf "Hydrodinamical methods of weather foreacast and climate investigation", St.Petersburg, Russia, 19–21 June, 2001 (in press, in Russian).
8. Boukhanovsky A.V., Kokh A.O., Rozhkov V.A., Savchuk O.P., Shaer I.S. Statistical analysis of water masses of Baltic sea. Proc. of GOIN, St.Petersburg, Gymiz, Russia, 2001 (in press, in Russian)
9. Cardone V.J., Cox A.T., Swail V.R. Specification of global wave climate: is this the final answer ?// 6th Int. Workshop on Wave Hindcasting and Forecasting. Monterey, California, November 6-10, p.211-223
10. Golube G.H., Van Loan C.F. Matrix computations (2nd ed.)., London, John Hopkins University Press, 1989.
11. Gould S.H. Variational methods for eigenvalue problems. University of Toronto Press, 1957.
12. Hamilton G.D. Processing of marine data. JCOMM Technical Report, WMO/TD #150, 1986.
13. Hansen I.S. Long–term 3–D modelling of stratification and nutrient cycling in the Baltic sea. Proc. of III BASYS Annual Science Conf., September 20–22, 1999, pp. 31–39.
14. Hotelling H. Relations between two sets of variables. Biometrika, 28, 1936, pp. 321–377.
15. Johnson R.A., Wichern D.W. Applied multivariate statistical analysis. Prentice-Hall Internalional, Inc., London, 1992, 642 pp.
16. Kalnay E., M. Kanamitsu, R. Kistler, W. Collins, D. Deaven, L. Gandin, M. Iredell, S. Saha, G. White, J. Woollen, Y. Zhu, A. Leetmaa, R. Reynolds, M. Chelliah, W. Ebisuzaki, W.Higgins, J. Janowiak, K. C. Mo, C. Ropelewski, J. Wang, R. Jenne, D. Joseph. The NCEP/NCAR 40-Year Reanalysis Project. Bulletin of the American Meteorological Society, №3, March, 1996.
17. Loeve M. Fonctions aleatories de second odre. C.R. Acad. Sci. 220, 1945.
18. Lopatoukhin L.J., Rozhkov V.A., Ryabinin V.E., Swail V.R., Boukhanovsky A.V., Degtyarev A.B. Estimation of extreme wave heights. JCOMM Technical Report, WMO/TD #1041, 2000.
19. Mardia K.V. Kent J.T., Bibby J.M. Multivariate analysis. London, Academia Press Inc., 1979
20. Ogorodnikov V.A., Prigarin S.M. Numerical modelling of random processes and fields: algorithms and applications. VSP, Utrecht, the Netherlands, 1996, 240 p.